Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2011, Article ID 158094, 8 pages
doi:10.1155/2011/158094

Research Article

Artificial Neural Networks for Classification in Metabolomic Studies of Whole Cells Using ^{1}H Nuclear Magnetic Resonance

D. F. Brougham,[1,2] G. Ivanova,[1,3] M. Gottschalk,[1] D. M. Collins,[1] A. J. Eustace,[1] R. O'Connor,[1,4] and J. Havel[5]

[1] *National Institute for Cellular Biotechnology, Dublin City University, Dublin 9, Ireland*
[2] *School of Chemical Sciences, Dublin City University, Dublin 9, Ireland*
[3] *REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[4] *School of Nursing, Dublin City University, Dublin 9, Ireland*
[5] *Department of Chemistry, Masaryk University, 611 37 Brno, Czech Republic*

Correspondence should be addressed to D. F. Brougham, dermot.brougham@dcu.ie

Received 9 April 2010; Revised 14 June 2010; Accepted 23 July 2010

Academic Editor: Mika Ala-Korpela

We report the successful classification, by artificial neural networks (ANNs), of ^{1}H NMR spectroscopic data recorded on whole-cell culture samples of four different lung carcinoma cell lines, which display different drug resistance patterns. The robustness of the approach was demonstrated by its ability to classify the cell line correctly in 100% of cases, despite the demonstrated presence of operator-induced sources of variation, and irrespective of which spectra are used for training and for validation. The study demonstrates the potential of ANN for lung carcinoma classification in realistic situations.

1. Introduction

Nuclear magnetic resonance spectroscopy (NMR, or MRS) has enormous potential for the study of biochemical and physiological changes in cancer tissues, due to its noninvasive nature and the large quantity of specific molecular information it can generate. Despite the sensitivity limitations of the technique, the inherent complexity of the spectra, and inevitable presence of overlapping resonances, there have been several successful NMR-metabonomics studies of cell tissue culture and culture extracts. The focus has been on elucidating the physiopathology of tumors and tumor cells, their drug toxicology and drug resistance, often with a view to identifying diagnostic markers [1–8]. A further significant complication in such studies arises from variability in the metabolite profile from sample to sample. This reflects many factors [9] including minor variations in growing conditions, the biochemical heterogeneity of the growing cells, the effect of different batches of sera (if used), and variations in cell and sample preparation. These additional factors may mask the inherent metabolite distribution, which may be diagnostic of the pathophysiological state of interest.

Experimental complications and difficulties also compromise the extraction of critical information from *in vivo* MRS experiments. In this case, the problems arise from the use of different MR-protocols, which affect the quality of the water suppression, differences in echo time and in the baseline, and so forth. While the causes are different in origin, they have a similar effect on the application. For both forms of magnetic resonance, many of these issues can, in principle, be addressed by improved experimental design, however, it is common for additional sources of variance to be identifiable only after extensive experimentation. In addition to technical issues are the natural physiological variability and the individual treatment history of the subject. As a result, there is an ongoing requirement for the development of magnetic resonance-based diagnostics using advanced statistical-, or other data-, analysis techniques which can reduce or compensate for additional sources of variability.

^{1}H NMR spectra of intact tissues or whole-cell samples are inherently complex due to the large number of contributing species which results in significantly overlapping resonance signals. Cell membranes also produce magnetic field inhomogeneity, further broadening the spectra [10]. In the case of cancer cells, a significant proportion of the lipids reside in a fluid environment and hence appear in the liquid-state ^{1}H spectra as strong "mobile-lipid" resonances [7, 8, 11]. Although the identification of the major resonances in ^{1}H NMR spectra can be used to characterise the metabolite profile, the complexity of the data sets usually necessitates the use of data reduction and pattern recognition techniques. These can provide information on the biochemical and physiological changes in cancer tissues, related to their physiopathology, drug toxicology, and drug resistance [12, 13]. Prominent amongst such techniques is principal component analysis (PCA), [14, 15] which involves diagonalisation of the spectral correlation or covariance matrix to identify independent sources of variance (principal components) across the set of spectra, and ranking of the components by their contribution to the overall variance. Thus, PCA is an unsupervised approach to data reprojection that can reveal the presence of classes, it has been applied to a variety of problems in biological science [16, 17].

Artificial Neural Networks (ANNs) belong to the so-called Artificial Intelligence group of methods, which were inspired by neurobiology and by the architecture of the human brain [18]. In recent times, these approaches have found applications in many branches of science. For example, they have been used in chemotaxonomy to classify limpets [19] from HPLC mass spectrometric data and in the identification of insect species from morphological measurements [20]. ANNs can be used to model data where the relations, or functions, are not known.

There have been some reports of the use of artificial intelligence and network methods in medical diagnostics which have involved analysis of magnetic resonance spectroscopic data. El-Deredy et al. [21] used ANNs to achieve reasonable prediction of the measured *in vitro* chemotherapeutic response from ^{1}H NMR of glioma biopsy extracts. More recently, Suna et al. [22] demonstrated the diagnostic potential of unsupervised approaches to classification by successfully analysing simulated ^{1}H NMR spectra using self-organising maps. This approach allowed the identification of stages along a metabolic pathway ranging from "normolipidaemic" to "metabolic syndrome". Tate and coworkers [23] reported the trial of an automated decision support system for classification of brain tumors from *in vivo* MRS, which showed a small but significant improvement in diagnostic accuracy over spectroscopy used and interpreted on its own.

In recent work [24], we reported PCA of ^{1}H NMR spectra recorded for a group of human lung carcinoma cell lines in culture and ^{1}H NMR analysis of extracts from the same samples. The samples studied were cells of lung tumor origin with differing chemotherapy drug resistance patterns. For whole-cell samples, it was found that the statistically significant causes of spectral variation were an increase in the choline and a decrease in the methylene and mobile lipid ^{1}H resonance intensities, which were correlated with our knowledge of the level of resistance displayed by the different cell lines. In this paper, we investigate the use of artificial neural network (ANN), a supervised method, to classify lung carcinoma. Two sets of whole-cell ^{1}H NMR spectra will be examined. These were recorded for two groups of human lung carcinoma cell lines, these were grown in culture and characterised over two different periods by two different groups of researchers (each consisting of a biologist and a spectroscopist), who both adhered to the same experimental protocol and used the same spectrometer. The cell lines studied include (i) the parent cell line DLKP, a human squamous nonsmall cell lung carcinoma; (ii) DLKP-A; (iii) DLKP-A5F, two resistant daughter lines; (iv) A549, a human lung adenocarcinoma cell line. The study also examines the capability of supervised techniques to compensate for experimental sources of variance, which may include operator bias and the cell culture growth process and in particular provide a test case for the application of ANN architectures in the identification and monitoring of resistance states in cancer tissue by MRS.

2. Experimental

2.1. Cell Samples. The cell lines DLKP [25, 26], DLKP-A [27], DLKP-A5F [28], and A549 were grown in culture to approximately 70–80% confluency in 175 cm^2 tissue culture flasks. Culture conditions were as follows: DLKP, DLKP-A, and DLKP-A5F and were cultured in minimal essential medium/Hams F12 (1 : 1, v/v) supplemented with 5% fetal calf serum and 2 mM L-glutamine. A549 was cultured in Dulbecco's modified Eagle's medium/Hams F12 (1 : 1, v/v) supplemented with 5% fetal calf serum. Cells were cultured as monolayers in tissue culture flasks and incubated at 37°C. A cell count was performed and c. 5×10^7 cells were separated and pelleted. These were then resuspended in deuterated PBS buffer and were kept in a container at 37°C before the start of the NMR measurements. The methods used were described in detail previously [24]. DLKP cells express a small amount of the multidrug resistance protein-1 (MRP-1) MDR drug efflux pump [25, 26]. DLKP-A [27] is a highly resistant clone of DLKP, which overexpresses the P-gp drug efflux pump. DLKP-A5F [28] was derived from DLKP by a different drug exposure profile, it is also highly drug resistant. A549 is an unrelated human lung adenocarcinoma cell line which was obtained from the American Type Culture Collection.

The first group of 13 samples, G1_13_21, were grown by a biologist during a six-month period, they were analysed by a first NMR spectroscopist. G1_13_21 contained 21 spectra and so was relatively sparse, it comprised DLKP [4 samples, 6 spectra], DLKP-A [4, 6], DLKP-A5F [3, 5], and A549 [2, 4]. The second group of 17 samples, G2_17_33, was grown independently, by a second biologist during a later six-month period and was analysed by a second spectroscopist [24]. G2_17_33 contained 33 spectra, it comprised DLKP [3, 6], DLKP-A [5, 10], DLKP-A5F [5, 9], and A549 [4, 8]. Thus for the integrated study presented here, a total of 30 samples were prepared and 54 ^{1}H spectra was recorded. The same protocols and methods were used by all the researchers

for cell growth and NMR spectroscopy. The biologist and spectroscopist who produced G1_13_21 will be collectively referred to as R1, and the biologist and spectroscopist who produced G2_17_33 will be referred to as R2. Due to the significant work involved in producing the large number of cells required for each spectrum, the number of samples in the study is inevitably somewhat limited. However, the total data set is larger than those usually reported in the analysis of NMR data by pattern recognition methods [16, 17, 29].

2.2. ^1H NMR Spectroscopy of Intact Cells. NMR spectra of the intact cell samples were recorded in deuterated PBS buffer on a Bruker DPX 400 spectrometer operating at 400.13 MHz for ^{1}H. Before all NMR experiments, the sample temperature was calibrated and controlled at 36.4 ± 0.2°C using an internal ethylene glycol thermometer (80% solution of ethane-1,2-diol in dimethyl sulfoxide-d^6). ^{1}H NMR spectra were acquired, without spinning, using WET [30] solvent suppression, with two Carr-Purcell-Meiboom-Gill (CPMG) echoes appended, using an echo delay of 1 ms [10]. Chemical shifts were referenced to an external 0.1% solution of sodium trimethylsilyl-[2,2,3,3-d$_4$]-propionate (TSP) in D_2O. All experiments were performed with a spectral width of 5200 Hz, an acquisition time of 3.15 s, and relaxation delay of 2 s. Three acquisition schemes were used to record the one-dimensional ^{1}H NMR spectra, all amounting to 128 scans. The first scheme (I) employed cycles of 16 dummy scans followed by four acquisition scans, $(16,4)_{32}$, giving an acquisition time of 3/4 hour. In the second scheme (II), 16 dummy scans were applied once prior to acquisition $16((0,16)_8)$, giving an acquisition time of 13 minutes. In the third scheme (III), 16 dummy scans and 128 acquisition scans were collected into 32 K data points, giving an acquisition time of 15 minutes. The time taken from resuspension to the start of data acquisition was typically less than 3/4 hour, and never more than 1 hour. All the data presented were recorded within 1 hour.

For the first group of 13 samples (G1_13_21) in the study, the acquisition schemes (I) and (II) were used for each sample. For the second group of 17 samples (G2_17_33), all three schemes were tested for each sample. Hence, the greater number of repeat spectra is for the second group. The inclusion of multiple spectra in the analysis from the same sample tests the stability of the samples over the time of the analysis. The insensitivity of the spectra to the sampling scheme used demonstrates that the samples do not change, for example, due to sedimentation, over the timescale that a single spectrum is acquired.

2.3. PCA Analysis. In the spectral region from 1.08 to 1.20 ppm, ethanol was observed, which was probably the result of endogenous processes. However, its intensity was highly variable, even within the same cell line, so this region was excluded from the analysis. The region containing the residual water resonance signal (3.56–6.05 ppm) was also excluded. The region above 6.05 ppm contained no features of sufficient intensity for reliable quantification, given the linewidth. For this study, we chose, as descriptors, the integrals over chemical shift regions (bins) of size 0.04 ppm

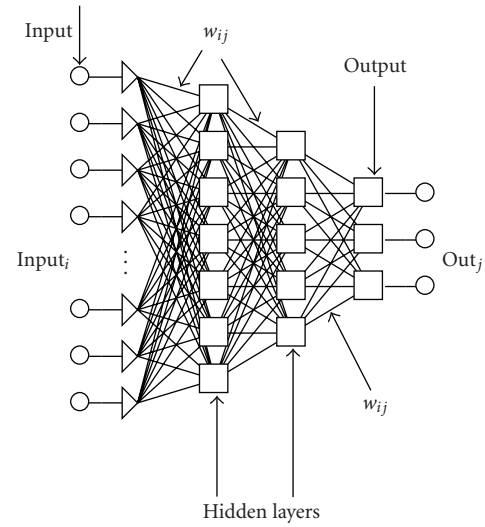

Figure 1: Schematic representation of a four-layer ANN architecture.

[12] which was found to produce the clearest separation of the cell types in the scores plots and the least noise in the corresponding loadings plots. Thus, the NMR spectra were reduced to 71 descriptors, with bin centres in the range 0.60–1.04, 1.24–3.56 ppm. We adopted the conventional approach [31] of normalisation relative to the total sum of the bin intensities in the region of interest. All the measures were implemented through writing an MATLAB (version 6.5.1, The Mathworks Inc.) code making use of the built in eigensolver.

2.4. ANN Analysis. ANNs are a sophisticated computational modelling tool, which can be used to solve a wide variety of complex problems. The attractiveness of ANNs comes from their capability to "learn" and/or model very complex systems and from the possibility of using them in classification. An ANN is a computational model formed from a certain number of single units, artificial neurons, or nodes, connected with coefficients (weights), w_{ij}, which constitute the neural structure. Many different neural network architectures can be used. One of the most common is the feed forward neural network of multilayer perceptions. The network is conventionally constructed with three or more layers, that is, input, output, and hidden layers, Figure 1.

Each layer has a different number of nodes. The input layer receives the information about the system (the nodes of this layer are simple distributive nodes, which do not alter the input value at all). The hidden layer processes the information initiated at the input, while the output layer is the observable response or behaviour. The inputs, input$_i$, multiplied by connection weights w_{ij} are first summed and then passed through a transfer function to produce the output, out$_i$. The determination of the appropriate number of hidden layers and number of hidden nodes in each layer is one of the most critical tasks in ANN design. Unlike the input and output layers, one starts with no prior knowledge of the number and size of hidden layers.

The use of ANN consists of two steps: "*Training*" and "Prediction". The "*Training*" consists first of selecting input and output data for the network. This data is referred to as the training set. In the training phase, where actual data must be used, the optimum structure, weight coefficients and biases of the network are identified. Training is considered complete when the neural networks achieve the desired statistical accuracy, that is, when they produce the required outputs for a given sequence of inputs. A good criterion to find the correct network structure and therefore to stop the learning process is to minimise the root mean square (RMS) error as follows:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j-1}^{M}\left(y_{ij} - \mathrm{out}_{ij}\right)^2}{N \times M}}, \tag{1}$$

where y_{ij} is the element of the matrix $(N \times M)$ for the training set or test set, and out_{ij} is the element of the output matrix $(N \times M)$ of the neural network, where N is the number of variables in the pattern, and M is the number of samples. RMS gives a single number, which summarises the overall error.

After a supervised network performs well on training data, it is important to check its performance with data that has not been used in training. This process is called *verification*. This testing is critical to insure that the network has not simply memorised the training set but has learned the general patterns involved within an application. At this stage, other input data are submitted to the network in order to evaluate if it can predict the outputs. In this case, the outputs are already known, but they are not shown to the network. The predicted value is compared to the experimental one to see how well the network is performing. If the system does not give reasonable outputs for this test set, the training period is not over or the network is able to model the data but cannot predict them.

In this work, ANN was used as a supervised method where a training data set was created from the library of NMR spectra, and the lung carcinoma classification of this training data set was known. The backpropagation method was used throughout. Firstly, the optimal ANN architecture was searched for and when the correct classification in the training phase was obtained, the usefulness of the created database and the prediction power of the networks were validated using an independent verification set. For the ANN analysis, we used 72 inputs; the 71 binned NMR intensities and the identity of the pairs of researchers (R1 and R2) as numbers 1 and 2. For output 4, nominal values were used, these identify the four cell lines, DLKP, DLKPA, DLKP-A5F, and A549, for which there were 12, 16, 14, and 12 spectra, respectively. All calculations were performed using the software Trajan Neural Network Simulator, Release 3.0 D. (Trajan Software Ltd 1996–1998, UK), on a standard PC computer running Microsoft Windows Professional XP 2000.

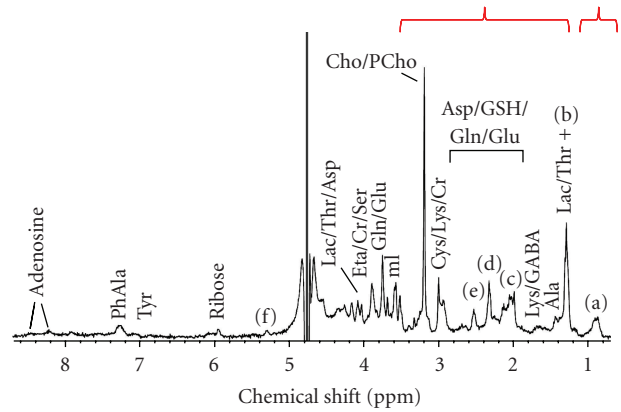

Figure 2: Typical 400 MHz ^{1}H NMR spectra of DLKP lung carcinoma whole cells. (a) CH_3, (b) CH_2, (c) $CH_2CH{=}CH$, (d) CH_2COO, (e) $=CHCH_2CH=$, (f) $HC{=}CH/CHOCOR$. The spectral regions used for statistical analysis (0.60–1.04 and 1.24–3.56 ppm) are indicated.

3. Results

3.1. ^{1}H NMR Spectroscopy of Whole Cells. A typical ^{1}H NMR spectrum of intact DLKP cells is shown in Figure 2. The appearance of the spectra and the assignment suggested below are broadly similar for all the cell samples analysed. A tentative assignment which is consistent with the literature [2, 4, 32, 33] is included in the figure [24]. Direct quantitative analysis of the whole-cell spectra is hampered by the potential multiple contributions from different metabolites to any given resonance line by the nonlorentzian lineshapes and by the broadness of the resonance lines. The resonances in the downfield region arise from species that are at low concentration, so quantification is precluded by the sensitivity limitations of the NMR measurement.

3.2. PCA Visualization of Whole-Cell Spectra. The binned NMR spectra of the intact cells were analysed using PCA. The scores plots are shown in Figure 3. Separation of the four cell types, within each of the two data sets, is apparent using the first two PCs, demonstrating that resistance type can be classified by PCA. It also demonstrates that the samples were stable over the course of the experiment and that the spectra are insensitive to the NMR sampling scheme. Loadings analysis shows that, for each data set, the spectral regions that contribute significantly to the first two principal components are from 1.24 to 1.50 ppm, corresponding to overlapped resonances from lipid methylenes and lactate methyls, and from 2.90 to 3.40 ppm, corresponding to overlapped resonances from N-methyl signals in the choline moieties of phosphatidylcholine, phosphocholine, and glycerophosphocholine. The contribution from other spectral regions to these two principal components is marginal.

Despite the fact that the same spectral regions allow separation within each data set, separation using PCA fails when the two sets of spectra are combined into one; see Supplementary Material available at doi:10.1155/2011/158094. It is apparent that, in addition to the metabolite differences

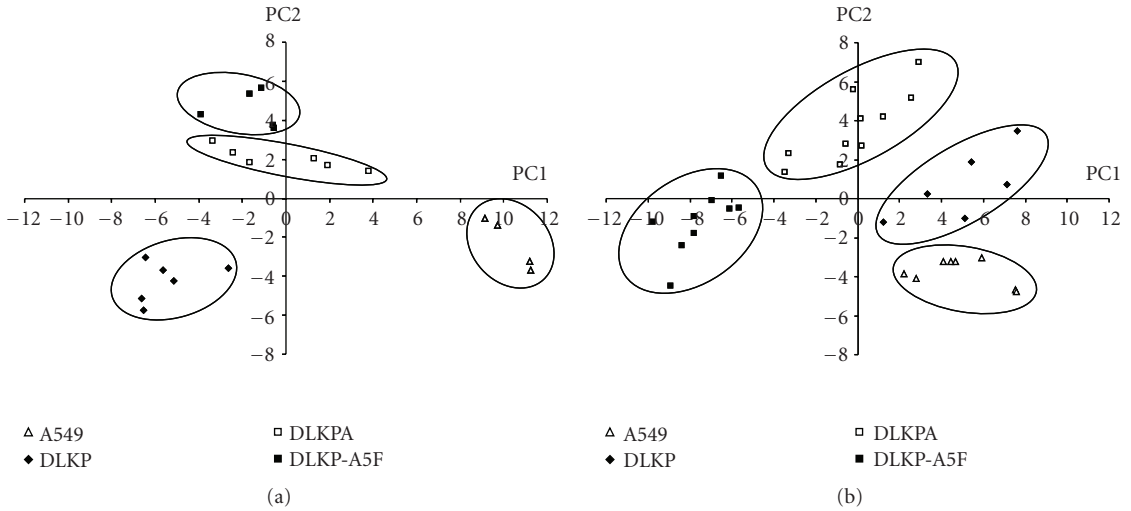

Figure 3: PCA scores plots for A549, DLKP, DLKPA, and DLKP-A5F, whole-cell data. Analysis is shown for G1_13_21 (a), G2_17_33 (b). The right hand panel is reproduced from [24] with permission.

of biological interest, there are subtle differences between G1_13_21 and G2_17_33 in the distribution of metabolites, which prevent classification of the entire (54 spectra) data set. The loadings analysis indicates contributions from across the spectral range, which may suggest variations in more than one metabolite. These spectral differences arise despite stringent efforts of the second group of researchers to adhere to the original experimental protocols and are reflected in the fact that there is not a simple correspondence between the orientation of the first two principal components between the two sets of spectra, Figure 3.

3.3. ANN Analysis of Whole-Cell Spectra. ANN analysis consists of separate training and verification steps. For this study, we adopted the strategy of choosing multiple verification sets of spectra at random from the 54 spectra available. In training, the first aim is to find an optimal ANN architecture to enable classification of the training data set. Several architectures of three up to four layered structures were examined for this purpose.

3.4. 3-Layers Architecture. Initially we adopted the simplest 3 layers architecture, in which case the search of the optimal architecture consists of optimising the number of nodes in the single hidden layer, effectively determining the corresponding weights, w_{ij}, to minimize the RMS (root mean square error) value according to (1). For our analysis, the RMS value ceases to decrease significantly above 5 to 6 nodes, Figure 4, we therefore used networks with 6 hidden nodes for verification. This optimal architecture will be labelled (72, 6, 4), with it we obtained an RMS = 1.38×10^{-3}. Figure 4 illustrates the process of searching for the optimal network architecture.

In spite of the fact that very low values for the residual mean squares were achieved using the (72, 6, 4) architecture, the appropriateness of the architecture and of the training set

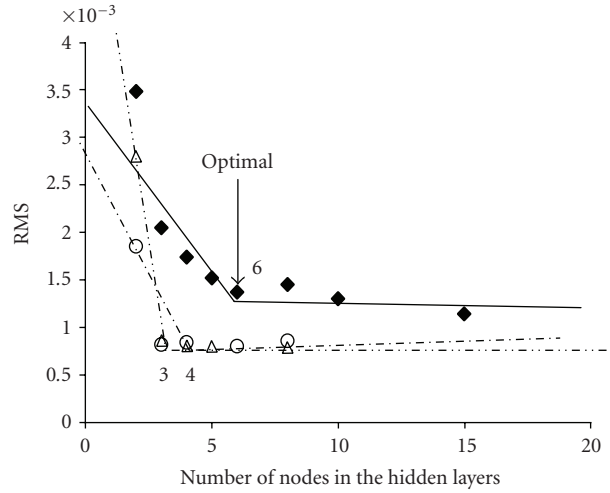

Figure 4: Plot of residual mean squares as a function of the number of nodes in the hidden layers, in the three-layers network (♦), and in the second (Δ) and third (○) layers of the four-layers network. For the networks labelled (Δ), 3 nodes were used in the third layer; and for the networks labelled (○), 4 nodes were used in the second layer. The lines have no physical meaning; they are included to better illustrate the optimal number of nodes.

was then tested with various verification sets, that is, a "cross-validation" procedure was undertaken. Initially, five spectra were randomly chosen and excluded from the training set and used then as the verification set. From 10 combinations and 10 independent networks trained, in only two cases were any of the 5 spectra classified as unknown, Table 1. These results are encouraging; two cases represent ~4% of the total, so for (72, 6, 4) the classification was verified as 96% successful. The failures may have arisen due to an insufficient number of spectra in the training set or because

Table 1: Results of cross-validation verification process for the three- and four-layer ANN networks.

Architecture (72, 6, 4)*		
Verification set no.	Spectra used in verification set	Results of Classification
1	2, 13, 17, 27, 38	all correct
2	21, 24, 31, 35, 51	all correct
3	4, 12, 22, 35, 44	spec. 35 classified as unknown
4	16, 17, 22, 25, 52	all correct
5	15, 16, 17, 23, 54	all correct
6	9, 15, 20, 24, 43	spec. 9 classified as unknown
7	3, 12, 15, 25, 51	all correct
8	19, 21, 43, 47, 54	all correct
9	16, 36, 37, 47, 48	all correct
10	12, 42, 44, 48, 50	all correct
Architecture (72, 4, 3, 4)		
1	5, 13, 20, 21, 22, 23, 24, 31, 51, 54	all correct
2	5, 8, 12, 15, 16, 29, 35, 36, 42, 49	all correct
3	8, 10, 13, 18, 23, 28, 33, 39, 40, 53	all correct
4	3, 5, 7, 9, 17, 27, 41, 45, 50, 52	all correct
5	5, 11, 16, 14, 20, 22, 24, 26, 44, 50	all correct

* where (72, 6, 4) refers to (the no. of inputs, the number of nodes in the hidden layer(s), the number of outputs).

networks with three layers have insufficient complexity for 100% prediction accuracy, in this case.

3.5. 4-Layers Architecture. We then examined networks with four layers (2 hidden). From several cases examined, it was found that four-layer ANN architectures performed similarly to simpler three layers architectures. Networks of the form (72, 4, 3, 4) or (72, 5, 4, 4) were investigated, note that the numbers in brackets refer to the number of inputs, the number of nodes in the first and in the second hidden layers, and the number of outputs. Acceptable RMS values, of 1.22×10^{-3} and 1.41×10^{-3} were obtained for (72, 4, 3, 4) and (72, 5, 4, 4), respectively, which are similar to the values obtained using the optimal three-layer architecture. Networks with the architecture (72, 4, 3, 4) performed very similarly to (72, 5, 4, 4) and require fewer unknowns (or weights, w_{ij}), 312 as opposed to 396. As a result, (72, 4, 3, 4) was found to converge faster and to be less sensitive to the number of spectra excluded from training to form the verification set. In fact, we found that 5 to 10 samples could be used for verification with 100% correct classification of the spectra, see Table 1. So in summary, the optimal 3- and 4-layer architectures were found to be (72, 6, 4) and (72, 4, 3, 4), respectively, Figure 5.

4. Discussion

The ^{1}H NMR spectra of intact cells for both G1_13_21 and G2_17_33 have similar general appearance with severe signal overlap and line broadening. Reprojection of either data set, using PCA, demonstrates that separation by cell types is possible due to systematic differences in the lipid methylene and lactate methyl resonances and the overlapped N-methyl ^{1}H nuclei of the choline-containing species [24]. Alterations in signal intensity and chemical shift from such cellular metabolites and biochemical intermediates have been described by other researchers in the area [6, 11]. However, because of the complex biochemical role played by these substances, we cannot ascribe a particular functional role to the findings, what is more the alterations appear to correlate and associate with particular phenotypic changes, for example, drug resistance. On the basis of the principal component analysis of either group, one could speculate that metabolite profiling by *in vivo* MRS has potential applications in monitoring the development of resistance in a given cancerous tissue. However, for the full data set such a possibility is effectively prevented by other influences on the metabolite distribution, which are comparable to, and nonorthogonal with, the "relevant" biochemical variation. We have shown that this significant obstacle can be eliminated, at least for *in vitro* studies of cell culture, by using a suitable ANN architecture. The most successful network was a four-layer structure with two hidden layers. After appropriate training, the (72, 4, 3, 4) architecture enabled 100% successful classification. Our approach may, in time, be expanded to the classification of larger data sets of spectra which have been recorded with less stringent control over sources of variance unrelated to the classification of interest. This result is encouraging and it is, to our knowledge, the first reported application of the use of ANNs specifically to correctly classify ^{1}H NMR spectra in a data set when additional "nonrelevant" sources of variance are included.

Other related examples of the combination of supervised and unsupervised methods include a report by Griffiths and coworkers [34], who obtained 85% accurate classification of meningiomas from nonmeningiomas, by initially using PCA to reduce the dimensionality of ^{1}H NMR spectra recorded for tumor biopsy extracts. The first thirty PCs

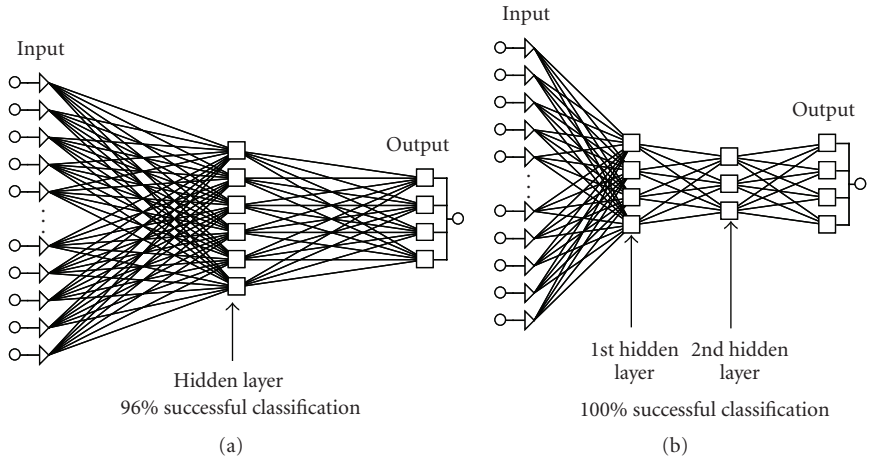

Figure 5: (a) Structure of the optimal 3-layer ANN architecture (72, 6, 4). (b) Structure of the optimal 4-layer ANN architecture (72, 3, 4).

from this first stage of analysis were then classified using a network. More recently, the performance of lineshape fitting and quantitative ANN analyses were compared by Hiltunen et al. [35] for both *in vivo* and simulated ^{1}H spectra. The good correlation obtained with these two approaches, for simulated data at least, suggested that ANNs have potential for quantification of *in vivo* MRS long echo time spectra. A further advantage of ANNs in the development of analysis methods for *in vivo* MRS is that they require less processing time than line fitting or other computational approaches [36]. Thus, our study adds to the growing number of applications of supervised techniques for exploiting the diagnostic potential of ^{1}H NMR spectra for biomedical purposes.

5. Conclusions

We have found that NMR data recorded for human lung carcinoma whole-cell culture samples can be used for analysis and classification. When sources of variation not directly related to the biological state of interest (drug resistance) are minimised or kept constant, visual separation of the cell type can be achieved using unsupervised pattern recognition techniques, such as PCA. On the other hand, when this condition is not met, in our case when different researchers were responsible for cell culture and spectroscopy, successful classification of the cell type could be achieved using artificial neural networks. The experimental and ANN methodology developed are a step towards the goal of robust and reliable diagnostics based on magnetic resonance spectral data. Furthermore, as similar experimental problems may be encountered in metabolomics applications using other spectroscopic techniques, biological classification using ANNs of data sets that include "nonbiological" sources of variance may be generally possible.

Acknowledgments

The authors acknowledge the support of the Higher Education Authority of Ireland, under the Programme for Research in Third Level Institutions (PRTLI3). D. Brougham, M. Gottschalk, and G. Ivanova acknowledge the financial support of the National Institute for Cellular Biotechnology, at DCU. They would like to thank the School of Chemical Sciences for its provision of spectrometer time. J. Havel would like to acknowledge the support of the EU Erasmus/Socrates exchange program between DCU and Masaryk University and to thank the Ministry of Education and Sports of the Czech Republic, Project LC 0635.

References

[1] U. Sharma, A. Mehta, V. Seenu, and N. R. Jagannathan, "Biochemical characterization of metastatic lymph nodes of breast cancer patients by in vitro ^{1}H magnetic resonance spectroscopy: a pilot study," *Magnetic Resonance Imaging*, vol. 22, no. 5, pp. 697–706, 2004.

[2] B. Martínez-Granados, D. Monleón, M. C. Martínez-Bisbal et al., "Metabolite identification in human liver needle biopsies by high-resolution magic angle spinning ^{1}H NMR spectroscopy," *NMR in Biomedicine*, vol. 19, no. 1, pp. 90–100, 2006.

[3] F.-G. Lehnhardt, C. Bock, G. Röhn, R.-I. Ernestus, and M. Hoehn, "Metabolic differences between primary and recurrent human brain tumors: a ^{1}H NMR Spectroscopic Investigation," *NMR in Biomedicine*, vol. 18, no. 6, pp. 371–382, 2005.

[4] M. C. Martínez-Bisbal, L. Martí-Bonmatí, J. Piquer et al., "^{1}H and ^{13}C HR-MAS spectroscopy of intact biopsy samples ex vivo and in vivo ^{1}H MRS study of human high grade gliomas," *NMR in Biomedicine*, vol. 17, no. 4, pp. 191–205, 2004.

[5] F.-G. Lehnhardt, G. Rhn, R.-I. Ernestus, M. Grne, and M. Hoehn, "^{1}H-and ^{31}P-MR spectroscopy of primary and recurrent human brain tumors in vitro: malignancy-characteristic profiles of water soluble and lipophilic spectral components," *NMR in Biomedicine*, vol. 14, no. 5, pp. 307–317, 2001.

[6] L. Le Moyec, O. Legrand, V. Larue et al., "Magnetic resonance spectroscopy of cellular lipid extracts from sensitive, resistant ,and reverting K562 cells and flow cytometry for investigating the P-glycoprotein function in resistance reversion," *NMR in Biomedicine*, vol. 13, no. 2, pp. 92–101, 2000.

[7] M. T. Santini, R. Romano, G. Rainaldi et al., "The relationship between ^{1}H-NMR mobile lipid intensity and cholesterol in

two human tumor multidrug resistant cell lines (MCF-7 and LoVo)," *Biochimica et Biophysica Acta*, vol. 1531, no. 1-2, pp. 111–131, 2001.

[8] A. Mannechez, B. Collet, L. Payen et al., "Differentiation of the P-gp and MRP1 multidrug resistance systems by mobile lipid ^{1}H-NMR spectroscopy and phosphatidylserine externalization," *Anticancer Research*, vol. 21, no. 6, pp. 3915–3919, 2001.

[9] J. L. Griffin, M. Bollard, J. K. Nicholson, and K. Bhakoo, "Spectral profiles of cultured neuronal and glial cells derived from HRMAS ^{1}H NMR spectroscopy," *NMR in Biomedicine*, vol. 15, no. 6, pp. 375–384, 2002.

[10] M. Bloom, K. T. Holmes, C. E. Mountford, and P. G. Williams, "Complete proton magnetic resonance in whole cells," *Journal of Magnetic Resonance*, vol. 69, no. 1, pp. 73–91, 1986.

[11] L. Le Moyec, R. Tatoud, M. Eugene et al., "Cell and membrane lipid analysis by proton magnetic resonance spectroscopy in five breast cancer cell lines," *British Journal of Cancer*, vol. 66, no. 4, pp. 623–628, 1992.

[12] M. Spraul, P. Neidig, U. Klauck et al., "Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 12, no. 10, pp. 1215–1225, 1994.

[13] J. K. Nicholson, J. C. Lindon, and E. Holmes, "'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," *Xenobiotica*, vol. 29, no. 11, pp. 1181–1189, 1999.

[14] J. C. Lindon, E. Holmes, and J. K. Nicholson, "Pattern recognition methods and applications in biomedical magnetic resonance," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 39, no. 1, pp. 1–40, 2001.

[15] E. Holmes, A. W. Nicholls, J. C. Lindon et al., "Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition," *NMR in Biomedicine*, vol. 11, no. 4-5, pp. 235–244, 1998.

[16] J. L. Griffin, K. K. Lehtimäki, P. K. Valonen et al., "Assignment of ^{1}H nuclear magnetic resonance visible polyunsaturated fatty acids in BT4C gliomas undergoing ganciclovir-thymidine kinase gene therapy-induced programmed cell death," *Cancer Research*, vol. 63, no. 12, pp. 3195–3201, 2003.

[17] Y. Wang, E. Holmes, J. K. Nicholson et al., "Metabonomic investigations in mice infected with Schistosoma mansoni: an approach for biomarker identification," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12676–12681, 2004.

[18] D. Zipser, "Identification models of the nervous system," *Neuroscience*, vol. 47, no. 4, pp. 853–862, 1992.

[19] J. Hernández-Borges, R. Corbella-Tena, M. A. Rodríguez-Delgado, F. J. García-Montelongo, and J. Havel, "Content of aliphatic hydrocarbons in limpets as a new way for classification of species using artificial neural networks," *Chemosphere*, vol. 54, no. 8, pp. 1059–1069, 2004.

[20] P. Fedor, I. Malenovský, J. Vaňhara, W. Sierka, and J. Havel, "Thrips (Thysanoptera) identification using artificial neural networks," *Bulletin of Entomological Research*, vol. 98, no. 5, pp. 437–447, 2008.

[21] W. El-Deredy, S. M. Ashmore, N. M. Branston, J. L. Darling, S. R. Williams, and D. G. T. Thomas, "Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks," *Cancer Research*, vol. 57, no. 19, pp. 4196–4199, 1997.

[22] T. Suna, A. Salminen, P. Soininen et al., "^{1}H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising maps," *NMR in Biomedicine*, vol. 20, no. 7, pp. 658–672, 2007.

[23] A. R. Tate, J. Underwood, D. M. Acosta et al., "Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra," *NMR in Biomedicine*, vol. 19, no. 4, pp. 411–434, 2006.

[24] M. Gottschalk, G. Ivanova, D. M. Collins, A. Eustace, R. O'Connor, and D. F. Brougham, "Metabolomic studies of human lung carcinoma cell lines using in vitro ^{1}H NMR of whole cells and cellular extracts," *NMR in Biomedicine*, vol. 21, no. 8, pp. 809–819, 2008.

[25] E. Law, U. Gilvarry, V. Lynch, B. Gregory, G. Grant, and M. Clynes, "Cytogenetic comparison of two poorly differentiated human lung squamous cell carcinoma lines," *Cancer Genetics and Cytogenetics*, vol. 59, no. 2, pp. 111–118, 1992.

[26] C. P. Duffy, C. J. Elliott, R. A. O'Connor et al., "Enhancement of chemotherapeutic drug toxicity to human tumour cells in vitro by a subset of non-steroidal anti-inflammatory drugs (NSAIDS)," *European Journal of Cancer*, vol. 34, no. 8, pp. 1250–1259, 1998.

[27] M. Heenan, L. O'Driscoll, I. Cleary, L. Connolly, and M. Clynes, "Isolation from a human MDR lung cell line of multiple clonal subpopulations which exhibit significantly different drug resistance," *International Journal of Cancer*, vol. 71, no. 5, pp. 907–915, 1997.

[28] C. O'Loughlin, M. Heenan, S. Coyle, and M. Clynes, "Altered cell cycle response of drug-resistant lung carcinoma cells to doxorubicin," *European Journal of Cancer*, vol. 36, no. 9, pp. 1149–1160, 2000.

[29] H. Hanaoka, Y. Yoshioka, I. Ito, K. Niitu, and N. Yasuda, "In vitro characterization of lung cancers by the use of ^{1}H nuclear magnetic resonance spectroscopy of tissue extracts and discriminant factor analysis," *Magnetic Resonance in Medicine*, vol. 29, no. 4, pp. 436–440, 1993.

[30] R. J. Ogg, R. B. Kingsley, and J. S. Taylor, "WET, a T_1- and B_1-insensitive water-suppression method for in vivo localized ^{1}H NMR spectroscopy," *Journal of Magnetic Resonance. Series B*, vol. 104, no. 1, pp. 1–10, 1994.

[31] J. E. Jackson, *A User's Guide to Principal Components*, Wiley-Interscience, New Jersey, NJ, USA, 2003.

[32] N. J. Waters, E. Holmes, C. J. Waterfield, R. D. Farrant, and J. K. Nicholson, "NMR and pattern recognition studies on liver extracts and intact livers from rats treated with α-naphthylisothiocyanate," *Biochemical Pharmacology*, vol. 64, no. 1, pp. 67–77, 2002.

[33] V. Govindaraju, K. Young, and A. A. Maudsley, "Proton NMR chemical shifts and coupling constants for brain metabolites," *NMR in Biomedicine*, vol. 13, no. 3, pp. 129–153, 2000.

[34] R. J. Maxwell, I. Martínez-Pérez, S. Cerdán et al., "Pattern recognition analysis of ^{1}H NMR spectra from perchloric acid extracts of human brain tumor biopsies," *Magnetic Resonance in Medicine*, vol. 39, no. 6, pp. 869–877, 1998.

[35] Y. Hiltunen, J. Kaartinen, J. Pulkkinen, A.-M. Häkkinen, N. Lundbom, and R. A. Kauppinen, "Quantification of human brain metabolites from in vivo ^{1}H NMR magnitude spectra using automated artificial neural network analysis," *Journal of Magnetic Resonance*, vol. 154, no. 1, pp. 1–5, 2002.

[36] H. Bhat, B. R. Sajja, and P. A. Narayana, "Fast quantification of proton magnetic resonance spectroscopic imaging with artificial neural networks," *Journal of Magnetic Resonance*, vol. 183, no. 1, pp. 110–122, 2006.